

Value production in a collaborative environment

Sociophysical studies of Wikipedia

Taha Yasseri · János Kertész

the date of receipt and acceptance should be inserted later

Abstract We review some recent endeavors and add some new results to characterize and understand underlying mechanisms in Wikipedia (WP), the paradigmatic example of collaborative value production. We analyzed the statistics of editorial activity in different languages and observe typical circadian and weekly patterns, which enabled us to estimate the geographical origins of contributions to WPs in languages spoken in several time zones. Using a recently introduced measure we showed that the editorial activity have intrinsic dependencies in the burstiness of events. A comparison of the English and Simple English WPs revealed important aspects of language complexity and showed how peer cooperation solved the task of enhancing readability. One of our focus issues was characterizing the conflicts or edit wars in WPs, which helped us to automatically filter out controversial pages. When studying the temporal evolution of the controversiality of such pages we identified typical patterns and classified conflicts accordingly. Our quantitative analysis provides the basis of modeling conflicts and their resolution in collaborative environments and contribute to the understanding of this issue, which becomes increasingly important with the development of information communication technology.

Keywords Peer-production · User-generated content · Wikipedia · Social dynamics · Burstiness · Human dynamics · Conflict · Language complexity

1 Introduction

Wikipedia is a truly amazing product of the 21st century. It is a free online encyclopedia¹ edited by volunteers, which has achieved within short period of time enormous success: This encyclopedia, which practically anyone can contribute to has a comparable reliability to the highly professional Encyclopedia Britannica [28] and has got by now the number one general work of reference in everyday practice. The main question related to Wikipedia is: How can an encyclopedia be reliable if anyone can edit it? The *bon mot* of Wikipedians is not a satisfactory answer, namely that “It works only in practice. In theory, it can never work.”

The literature about Wikipedia (WP) is overwhelming. Without seeking completeness, Okoli et al. [67] tracked more than 2000 related articles. However, there are rather comprehensive reviews, e.g.,

Supported by EU’s FP7 FET Open STREP Project: ICTeCollective No. 238597.

T. Yasseri

Department of Theoretical Physics, Budapest University of Technology and Economics, Budapest, Hungary.

E-mail: yasseri@phy.bme.hu

J. Kertész

Center for Network Science, Central European University, Budapest, Hungary,

and Department of Theoretical Physics, Budapest University of Technology and Economics, Budapest, Hungary.

E-mail: janos.kertesz@gmail.com

¹ <http://en.wikipedia.org/wiki/Wikipedia>

[66,44] and an overview of the visibility of WP in scholarly publications [69]. In addition, there are also online platforms to collect and index WP-related academic literature; among them are “WikiLit”² [7], “AcaWiki”³ and “WikiPapers”⁴. A monthly review of the most recent scholarly studies on WP is also available at “Wikimedia research Newsletter”⁵.

First Wikipedia studies were mostly on its size and growth, showing an initial exponential growth [99,4], which later reported to be saturating by other authors [90]. Another main line of WP research is focused on vandalism detection [86,72,107,102,3]. Assessing user reputation [2,42] and investigating the articles quality [39,105,61,88,43,51,41,106] are other two important topics. To understand the management system of WP, there have been interesting studies on user authority, adminship, governance and promotion strategy [59,1,24,83,58], in addition to analysis of WP policies and bureaucracy [13]. A considerable amount of WP policies are on what to be/not to be in WP. Consequently, there are studies on topical coverage and notability of entries [38,33,93]. Seeing WP as a network of articles, various researchers offer analysis and models for topology and growth of the Wikigraph [73,15,12,17,16], whereas some others used WP to build up knowledge taxonomies and semantic structures [87,64,71,14,113,50,85,89]. Masucci et al. showed that semantic space has a scale-free structure by analyzing information extracted from WP [63]. More to the sociological side, Restivo and van de Rijt studied the effect of social awards on users activity [77] and Lam et al. explored the gender imbalance among WP editors [54]. Massa presented an algorithm to extract the social network of editors [62], and Danescu-Niculescu-Mizil et al. have studied talk page conversations to observe the relation between language coordination and social power of editors [21]. Clearly, scholarly studies in the field of peer-production go beyond WP and for instance Roth et al. studied dynamics of communities of wiki-based projects in the whole “WikiSphere” [78].

Our motivation to study Wikipedia comes from the need of understanding the laws of modern collaborative value production. This is of great importance as in our increasingly complex world the role of information communication technology (ICT) mediated peer collaboration is expected to become more and more important in the future. Due to its relation to ICT, the methods of “Computational Social Science” (CSS) [57] are adequate tools of investigation of such collaborations. CSS is a truly multidisciplinary endeavor with a considerable contributions from physicists (see, e.g., [18]). The main difference between traditional social science and CSS is that the latter is data driven: it uses the digital footprints we leave behind in almost all our activities in the digital era [10].

Collaboration has always been fundamental to most human achievements. Modern information communication technology opens up entirely new ways of cooperation, where partners can interact remotely with an unprecedented speed exchanging extremely large amount of information. Tim Berners-Lee developed originally the World Wide Web [103] at CERN in order to create an appropriate platform for huge collaborations, which are ubiquitous in high energy physics. Another important example is that of free software development as defined by the Free Software Foundation⁶. Nowadays all major scientific projects from the Human Genome⁷ to Hubble Space Telescope⁸ rely heavily on ICT mediated collaboration but even on smaller scale we often use Current Version Management⁹, wiki [60] and related environments to increase efficiency. WP is a paradigmatic example of collaborative environment with the additional advantage that all the changes and interactions are well documented and publicly available, which makes it particularly suitable for scientific studies.

Many questions arise when studying WP from our point of view. What are the characteristic features of editorial activities? How are they related to other examples of human dynamics, which have been intensively studied in CSS [8,48]? What is the mechanism behind the emergence of an article? How can the complexity of the product of the cooperation, namely that of the articles be characterized?

² http://wikilit.referata.com/wiki/Main_Page

³ <http://acawiki.org/Home>

⁴ http://wikipapers.referata.com/wiki/Main_Page

⁵ <http://meta.wikimedia.org/wiki/Research:Newsletter>

⁶ <http://www.fsf.org/>

⁷ http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml

⁸ http://www.nasa.gov/mission_pages/hubble/story/the_story_2.html

⁹ <http://savannah.nongnu.org/projects/cvs>

Table 1 Article statistics for 10 largest Wikipedias. First and second columns are indicating the Language and the Symbol of the Wikipedia editions. In the following columns number of Articles (divided by 1000), Average Length of the articles in characters, number of editors with at least one edit, divided by the number of articles, and number of Featured articles are reported.

Language	Symbol	Art. (k)	Av. Len.	Av. Edit/Art.	Editor/Art.	Featured
English	en	4,080	5544	136.7	1.26	3638
German	de	1,454	5081	77.3	0.39	2113
French	fr	1,287	5189	68.0	0.32	1093
Dutch	nl	1,072	2567	30.6	0.14	272
Italian	it	957	4799	59.5	0.20	538
Polish	po	917	3634	35.6	0.15	530
Spanish	es	915	5027	69.2	0.53	1035
Russian	ru	892	7913	59.1	0.24	553
Japanese	ja	826	6357	54.6	0.32	66
Portuguese	pt	739	3421	43.5	0.26	694

How do conflicts emerge and get resolved? In the following we will present analysis of WP data in order to contribute to the clarification of these questions.

To accustom the unfamiliar readers to the terminology and work-flow of Wikipedia, in the next Section we briefly review main tools and objects in Wikimedia platform. Familiar readers are encouraged to skip this Section. In Section 3, we explain different methods and sources for collecting WP data, and in Section 4, a summary of our recent results [91,92,110,111,109,94] is provided and compared to the related reports by other authors. We close the paper with a conclusion (Section 5).

2 How Wikipedia works

WP has more than 280 language editions at the moment. Main concepts and structures are similar in all language editions with little variations due to local modifications by the editors' community of the specific edition. Later we will deal with several WPs, however, whenever it is not specified else explicitly, the English WP is meant.

Describing the structure of WP, there are two main elements to name, i) Articles ii) WP editors, also called "Wikipedians". The rest is all about the internal and inter-element connections and interactions of the members of these two groups, which we name "Accessories" in this manuscript.

2.1 Articles

Wikipedia, similarly to almost any other encyclopedia consists of entries about different topics, hereafter called Articles. Each article of WP has necessarily a "title", a nonempty "content", and a "history" which is a collection of all previous revisions of the article beginning from its inception. In Table 1, some basic statistics of articles in the ten largest WPs are given. Articles of each language edition are connected via internal links. That makes the whole language edition a directed graph. Ideally this graph should be connected, however there are always "Orphan" (not linked by any article) and "Dead-end" (not linking to any other article) articles. There are also inter-language links, connecting articles from different language editions.

In general, articles could be edited by any Internet user. However there are protections against vandalism applied to some articles and prohibiting different class of editors from editing. Access to more complex actions, e.g. creating a new article, changing the title, or deleting an article is also subject to hierarchal structure of editors (described in the next Section).

Featured articles: "Featured articles are considered to be the best articles Wikipedia has to offer, as determined by Wikipedia's editors."¹⁰ Articles are tagged as featured based on the community decision

¹⁰ http://en.wikipedia.org/wiki/Wikipedia:Featured_articles

Table 2 Editor statistics for 10 largest Wikipedias. First and second columns are indicating the Language and the Symbol of the Wikipedia editions. In the following columns, number of Registered users, users who have actually Contributed (at least one edit), Administrators, Bureaucrats, and the editors who are banned forever, are reported.

Language	Symbol	Registered	Contributed	Admins	Bureaucrats	Banned
English	en	17,186,079	5,085,719	1,461	34	93,812
German	de	1,467,633	564,993	268	5	7,978
French	fr	1,332,309	413,091	195	7	3,214
Dutch	nl	469,358	147,758	64	9	1,011
Italian	it	773,446	192,198	105	6	2,995
Polish	po	501,381	138,804	157	6	731
Spanish	es	2,292,694	485,802	134	133	5,473
Russian	ru	886,133	215,759	92	5	2,855
Japanese	ja	643,770	260,206	61	9	5,693
Portuguese	pt	1,026,749	198,000	37	7	1,150

on their accuracy, neutrality, completeness and style. In English WP there are more than 3,500 featured articles (see Table 1).

Lists of controversial articles: There are also lists of articles with severe editorial disagreements in their history. For example “List of Controversial Articles”¹¹, and List of “Lamest Edit Wars”¹². However, the accuracy and coverage of those lists are questionable. There is no clear definition and systematic algorithm to determine, which articles should be listed.

2.2 Wikipedians

In principle any person with access to Internet could be a Wikipedia editor. Editors are recognized by the system based on the IP addresses, through which they are connected or with their user-name which is assigned upon registration. As long as editors edit via their user-names, in general no personal information about them is revealed, unless voluntary disclosure by themselves. There are semi-annual surveys run by Wikimedia Foundation to provide some demographical information about the community of WP editors.¹³ However, since participation in the survey is completely voluntary, the reliability and coverage of this information is questionable. Therefore, personal information of the editors community of WP, is the most unknown aspect of it.

There is a well defined hierarchal structure among Wikipedians, such that editors from different classes have access to certain editorial actions. For exact description of each level rights and accessed, see http://en.wikipedia.org/wiki/Wikipedia:User_access_levels. In brief, some of these classes, common in all language editions are: a) Unregistered users, with the right to edit unprotected existing pages. b) New users, with the right to edit unprotected pages and create new pages. c) Auto-confirmed users, with the right to edit semi-protected pages and move pages to new titles. d) Administrators (admins), with the right to edit protected pages, delete or protect pages, and block other editors from editing, c) Bureaucrats, with the right to change the user rights and in most of the Wikipedias, conclude the promotion polls. Promotion to higher levels, starting from adminship, is upon consensus of editors community confirmed by promotion polls. In Table. 2, some basic statistics on the editor communities of 10 largest language editions are given.

2.3 Accessories

Around half of the edits in Wikipedia are outside the main name-space, i.e. in accessory pages [52]. This pages control the underlying mechanism of growth and maintenance of WP articles in the main name-space. Here we briefly describe some of them.

¹¹ http://en.wikipedia.org/wiki/List_of_controversial_articles

¹² http://en.wikipedia.org/wiki/Wikipedia:Lamest_edit_wars

¹³ http://meta.wikimedia.org/wiki/Editor_Survey_2011

Table 3 Page statistics for 10 largest Wikipedias. First and second columns are indicating the Language and the Symbol of the Wikipedia editions. In the following columns, number of All pages, Articles, Article Talk pages, User Pages, User Talk pages, and Categories, and in the last column sum of number of Wikipedia guidelines, projects, polls and Help pages are reported. All the numbers are divided by 1000.

Language	Sym.	All	Art.	Art. Talk	Us. Page	Us. Talk	Cat.	WP, Help
English	en	28,068	4,080	4,212	1,527	8,057	891	737
German	de	4,062	1,454	458	338	342	153	35
French	fr	5,268	1,287	1,024	166	951	213	33
Dutch	nl	2,320	1,072	73	107	456	70	15
Italian	it	3,045	957	188	64	851	172	96
Polish	pl	1,766	917	219	73	95	101	26
Spanish	es	3,845	915	184	125	971	184	23
Russian	ru	3,053	892	362	77	245	212	28
Japanese	ja	2,232	826	163	75	332	100	77
Portuguese	pt	2,994	739	378	70	922	145	58

Policies, guidelines, essays and instructions: “Wikipedia’s policies and guidelines are pages that serve to document the good practices that are accepted in the Wikipedia community.”¹⁴ These policies are however subject of change and improvement by the community of editors and may slightly differ among different language editions.

User pages: “User pages are for communication and collaboration.”¹⁵ They could be used to provide personal information of the editor or less encyclopedic content related to the editor. However, as they are part of the encyclopedia project, their content should not violate the main guidelines.

Article talk pages: The purpose of a Wikipedia talk page is to provide space for “editors to discuss changes to its associated article or project page”.¹⁶ Talk pages are the main channels for social interactions between articles, and supposed to be the main place to resolve disagreements and editorial conflicts.

User talk pages: User talk pages are designed for more general communications directly to each editors. User talk pages are usually less technical than article talk pages and conversations are more personal.

Common discussion pages: apart from article and user talk pages, there other discussion pages related to specific projects, polls, and more collective activities. There are also different communication channels for Wikipedians outside of the WP, e.g., IRC channels and Wikimedia mailing lists; for an overview see [70].

Categories: Categories are intended to group together pages on similar subjects.¹⁷ Categories are a feature of MediaWiki platform. It allows articles to be grouped in categories and provides the facility for the readers to navigate through the related articles. The process of article categorization, is carried out by editors, and its accuracy is at the same level of other content of WP.

3 Methods and Data

Beyond usual statistical methods to study Wikipedia, there are numerous open source software packages for different analyzing tasks. Among them is “WikiTrust”¹⁸ [2], to measure article quality and assign a

¹⁴ http://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines

¹⁵ http://en.wikipedia.org/wiki/Wikipedia:User_pages

¹⁶ http://en.wikipedia.org/wiki/Wikipedia:Talk_pages

¹⁷ <http://en.wikipedia.org/wiki/Wikipedia:Categorization>

¹⁸ <http://www.wikitrust.net/>

reputation to it. WikiXray¹⁹ [25] is another package for doing in-depth statistical analysis on different parameters, e.g. size of WPs, size of articles, number of contributors to each article, etc. However, since all WP data is publicly available, developing home made packages to analyze this data is a common approach.

3.1 Data

Every single action of Wikipedia editors is tracked and recorded. This includes all edits on articles, posts on talk pages, page deletions or creations, changes in page titles, uploading multimedia files, etc. Apart from the practical advantages of this complete archiving, it is also extremely valuable from scientific point of view. WP is one of the few human societies that the history of all actions of its members are recorded and accessible.²⁰

Live data: There are two convenient ways to access live data of Wikipedia. i) “Wikimedia Toolserver”²¹ databases, which contains a replica of all Wikimedia wiki databases, and ii) “MediaWiki web service API”²². For statistical analysis of contributions, Toolserver database tables are among the best sources of information.

Dumped data Wikipedia also offers archived copies of its content in different formats²³, e.g., XML and HTML and different types, e.g., snapshots of full history of articles or a collection of latest version of all articles. Generally for historical text analysis of articles, the most reliable source would be these static copies.

Semantic Wikipedia “Semantic Wikipedia”, as a general concept would be a combination of Semantic Web and WP data to provide structured data sets through query services. There are various projects providing access to Semantic WP. Examples are “DBpedia”²⁴ [6], “Semantic MediaWiki”²⁵ [98], and “Wikipedia XML corpus”²⁶ [23]. For a list of Semantic WP projects see http://en.wikipedia.org/wiki/Wikipedia:Semantic_Wikipedia.

4 Results and Discussion

4.1 Editorial habits

Similarly to any other large human society, the community of Wikipedia editors is very inhomogeneous. Editors vary in age, gender, nationality, education, occupation, religion, interest, etc.

4.1.1 Edits statistics

Heterogeneity is present in the level of activity. Not only the total number of edits by each editor has a largely extended distribution [104, 68, 40], but also the number of different articles each editor contributes to is varying considerably from one editor to another [25]. Finally, the number of editors contributing to an article has also a fat-tailed distribution (see Fig. 1), however with a lower cut-off when we only consider a selection of “Featured Articles”, which are supposed to be articles with high level of completeness and accuracy.

¹⁹ <http://meta.wikimedia.org/wiki/WikiXRay>

²⁰ Except deleted revisions, which are only available for admins and higher, and “overseen” revisions, which are accessible by no one.

²¹ <http://toolserver.org>

²² <https://www.mediawiki.org/wiki/API>

²³ <http://dumps.wikimedia.org>

²⁴ <http://dbpedia.org>

²⁵ <http://semantic-mediawiki.org>

²⁶ <http://www-connex.lip6.fr/~denoyer/wikipediaXML>

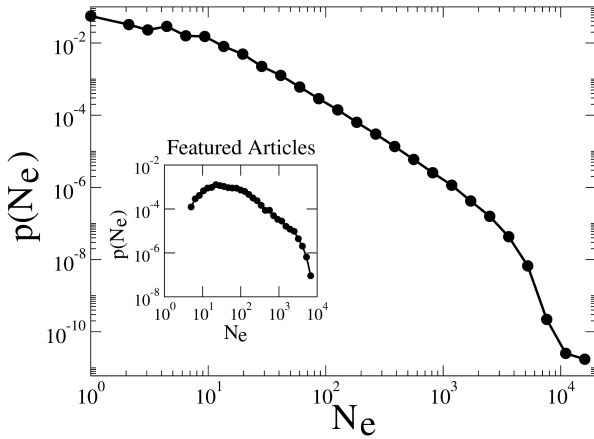


Fig. 1 Probability distribution function of number of different human-editors contributed to each article. Most of the articles are basically edited by few editors, and few of them are edited by a large editorial pool of 10,000 editors. *Inset:* Probability distribution function of number of editors of the 3,122 “featured articles”. The lower cut off of the distribution moves from its natural value of 1 for the whole sample, to a value of 5 editors at least in the featured articles sample. The peak of the distribution is about 22 editors, which could be considered as the optimal number of editors to achieve an acceptable quality for an article.

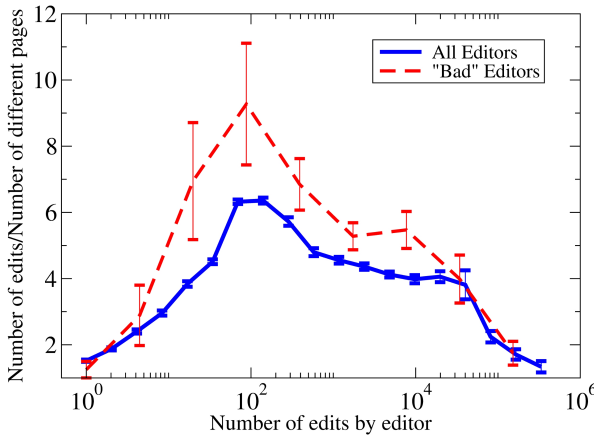


Fig. 2 Average number of edits per article vs. total number of edits by each editor, for all editors (blue) and a group of “Bad Editors” with large number of penalties during their editorial life (red). For both samples, the average value increases initially, meaning a trend towards more concentration on limited number of pages, followed by a decrease signaling broadening of field of interest and editorial zone of editors as they get more experienced. The consenter at the peak is more intense for Bad Editors and their contribution is focused on fewer pages compared to average editors.

One source of the inhomogeneity is that statistical characteristics of editors show time evolution; different phases can be identified in the editorial behavior as they get more and more mature. For example, in Fig. 2 the number of edits per different articles as a measure of the editors’ focus, versus total number of edits, calculated for a large sample of editors (all editors with at least 1 edit) is showed. The general trend is as follows: Editors start by less intensive edits on different pages, then gradually they get more focused on few articles and the ratio of the number of edits per number of different articles increases. Once they reach a level of maturity, again their field of interest becomes wider and finally extremely senior editors, distribute they editorial efforts on a huge number of articles. Note that this is the average trend and again may differ from editor to editor. In the same plot, same quantity is shown for a sample of “Bad Editors” who are blocked from editing at least 7 times per 1000 edits. Although the overall trend is the same as the whole sample, but a larger peak, indicating more intensive focus on few pages, is clearly visible. This is very intuitive: Editors seeking conflict and having the tendency to violate the guidelines have special interests in a limited number of pages, where they disturb the collaborative environment.

4.1.2 Time of editing

Since all edits are recorded along with a timestamp, it is very convenient to perform temporal analysis on editorial activities at different time scales.

Burstiness: Most of the editors do their edits following a certain type of inhomogeneous temporal pattern. There are periods or bursts of high activity separated by low or no activity intervals. Compared to a homogeneous Poisson process the distribution of the inter-event times has a much fatter tail in the case of a bursty pattern. One trivial source of the temporal inhomogeneity is the circadian

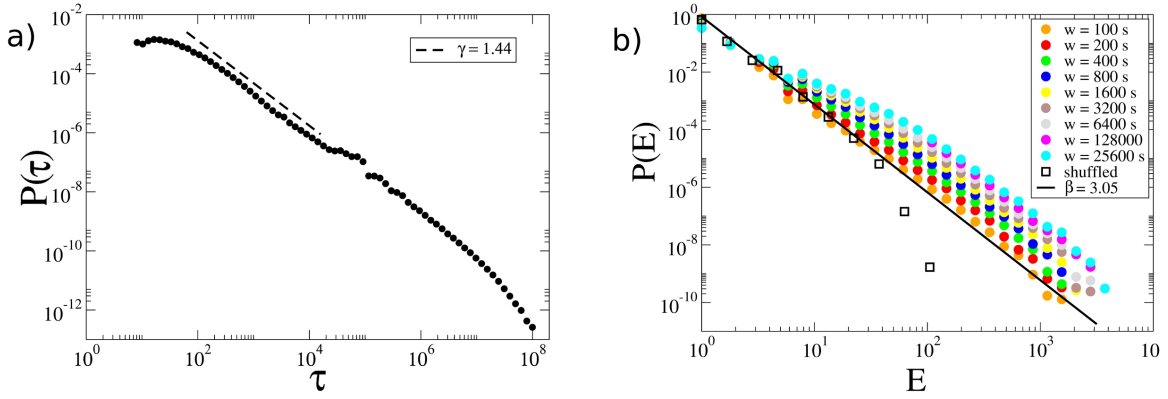


Fig. 3 Two characterizing functions of the temporal pattern of editorial activity at individual editors level averaged over a sample of 100 most active editors. a) Probability distribution function of the inter-edit time intervals (in seconds) fitted with a power law of the exponent $\gamma = 1.44$. The bump is due to the circadian pattern and corresponds to 24 hours. b) Probability distribution function of the number of edits in the bursty periods separated by windows of silence with the width of at least w . Color circles are the original data and empty squares corresponds to the shuffled sequence. The exponent of the power-law fit is $\beta = 3.05$ and the decay for the shuffled data tends to an exponential form. From panels (b) it becomes clear that there is long term correlations in edit trains at the level of editors in addition to broad distribution of time intervals shown in panel (a). *These figures were originally published in [111] under the terms of the Creative Commons Attribution License.*

pattern of human activity. However, recently it was shown by introducing the notion of bursty periods that “burstiness” often originates from memory effects and reinforcement mechanisms [48]. These bursty periods are separated by intervals of length w of no activity and the distribution of the number of events in the bursty periods follows a power law in contrast to the memory free case, where it is exponential. Our investigations on short time scale temporal features of editorial activities, reveal strong evidences for the presence of similar mechanisms [111]. For instance, in Fig. 3 two characterizing measures of temporal patterns for the activity train of a class of active users are shown. First, the distribution of inter-edit time intervals (Fig. 3 (a)) is extremely fat-tailed, indicating the presence of long silence periods. Second, the distribution of the number of edits in bursty periods follows a power law, which is not sensitive to the choice of w (Fig. 3 (b)). However, in the absence of memory effects, these the latter would be an exponential distribution [48], as it is indeed the case when we shuffle the data Fig. 3 (b).

Ung and Dalle, also reported a power law distribution of the inter-edits time intervals and interpreted their observation as an outcome of editors’ focus on few certain tasks (articles). They measured the slope for different class of users and show that more/less skewed distributions correspond to more focused/dispersed editors [96].

Daily patterns: As mentioned above, the activity pattern of individual editors are quite heterogeneous in time. However, if we consider the whole editorial pool of a language edition of Wikipedia, we can define an average activity level for all editors which also has its own large scale characteristics. In [110], it is shown that WP is mostly edited between 1 pm and 11 pm, almost in a universal manner for all language editions. This is in accord with the results in [76, 47]. Deviations from this universality originates from cultural differences and working habits, such that language editions with more editors from countries with longer working hours, are even more edited in later time in evening and around midnight. In addition, for more globalized language editions, the activity curves are flattened, due to contributions from different time zones (see 4.1.3).

Weekly patterns: Weekly patterns are quite universal within one WP and different WPs can be classified in different categories based on the activity pattern of their editors [110]. For example, German, English, Spanish, and Italian WPs are mostly edited during the working days, in contrast to Japanese, Korean, and Chinese WPs being mostly edited on weekends. Our findings are in accord with [76] but in contrast with in [47]. However, the latter work studied a sample of four languages only and a shorter

monitoring time, and we believe that these made them conclude that editorial activity in WP “while showing a clear diurnal pattern, do not have a clear weekday-weekend pattern.”

4.1.3 Edits origin

As mentioned earlier in Section. 2.2, personal information of editors is rarely available. That includes their nationality and living place. However, to understand many aspects of social characteristics of the editors societies, as well as conflicts and potential biases in content, such information could be crucial. To achieve exact data on the location of editors, analysis must be restricted to unregistered users with edits recorded along with IP addresses, whose edits are typically between 5% to 10% of the total community contributions in different language editions, and clearly not representing the whole community. Moreover, a considerable part of such editors are atypical (vandals, single act editors). Nevertheless, Hardy et al. followed edits of 2.8 Million such editors and geolocate the them and the edited articles. By counting the number of edits as a function of distance between editor and article, an exponentially decaying distance dependence was obtained [34]. Cohen has investigated the contribution of unregistered editors to English WP and concluded that most of unregistered edits are from large cities and metropolitan areas [19]. However, normalization to population of regions seems to be a missing essential for such conclusion.

Based on the results on daily patterns of editing for geographically localized WPs, such as German, Italian, Hungarian and defining a “standard activity pattern”, one could estimate the global distribution of editors to globalized language WPs in the following way; initially some candidate regions, from which large population of editors would contribute to the given WP are selected. In the next step, a linearly weighted superposition of standard activity patterns shifted to the local time of the candidate regions is made. By minimizing the difference between this composed activity pattern and the activity pattern constructed from the real data, a set of optimal weights is obtained. Clearly, these weights are proportional to the share of editors contributing from the corresponding regions. Surprisingly, it turned out that English WP is almost equally edited by North Americans and editors from the rest of the world [110]. In Fig. 4, estimations for the share of contributions from different regions to each language edition is shown.

4.1.4 Characterization of edits

As mentioned above, each editor has her own unique characteristics and editorial personality. However, similar patterns could be observed by considering types of edits. In a novel approach, Wettenberg et al. established a visualization method to illustrate different editorial actions, e.g. adding, spelling correction, reverting, etc. in a time sequence. In the next step based on the patterns of activity they could distinguish different kinds, namely systematic activity, reactive and mixture activity patterns [101].

Kittur et al. have classified editors based on number of edits and also specifically followed admins’ contributions from the inception of WP [52]. They concluded that in the beginning of the WP history, large amount of contributions were offered by “elite users, however, it has gradually changed in a way that after 2004, average users overtook the elites. By counting the number of added and removed words for different editors, they suggest that elite users in average add more words per edit compared to normal editors.

We considered the volume of contributions by measuring the volume of each edit in unit of characters. Naturally, negative volume is assigned to deletion. In Fig. 5 two examples of edit volume profile of two different type of users is shown: Fig. 5 (a) shows a typical “producer” and Fig. 5 (b) shows a typical “maintainer”. It is needless to mention that there are many different type of users, but by analyzing the edit volume profile of few tens of most active editors, these two types are found dominantly. In the other words, although the results shown here are for two single editors, they are representative of large groups of editors, who can be categorized in one of the these groups based on their edit volume profile. Note that, for the case of producer editors, a separation between additions in the size of “sentences (first peak) or “paragraphs” (second peak) is nicely visible.

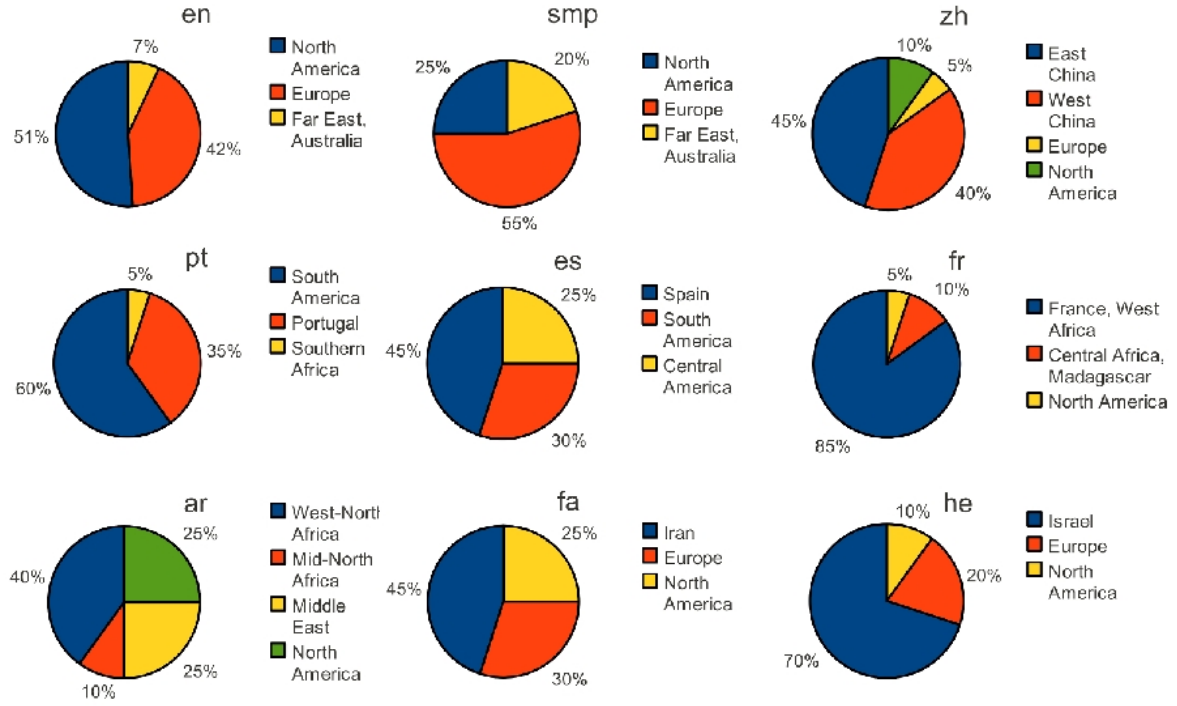


Fig. 4 Estimations for the share of editors from different regions to different Wikipedia language editions. In the first row, *en*, *smp* and *zh* stand for English, Simple English and Chinese WPs. The share of North America to the English WP hardly goes beyond half and it is around one quarter for Simple English WP. In the middle row, *pt*, *es*, and *fr* stand for Portuguese, Spanish, and French. The low level of contributions to French WP from North America (Canada) is worth mentioning. In the lower row, three Middle East languages are shown; *ar*, *fa*, and *he* stand for Arabic, Persian, and Hebrew. Here, large amount of contributions from western regions to Persian WP is notable. *This figure is originally published in [110] under the terms of the Creative Commons Attribution License.*

4.1.5 Linguistic features

The content of Wikipedia is generated by large number of editors collaboratively and without any professional or external supervision. That makes the resulting written language of WP articles a unique multilingual corpus of natural languages. A single sentence in WP might be written, edited and polished by various editors many times, therefore any personalization bias is eliminated on large scales. Moreover, the fully recorded history of articles give the opportunity to follow the short time scale evolution of language and characterize the gradual changes of written language in the digital era. Finally, since WP is huge, and available in many different languages, statistical approaches can be taken in a proper way.

In a practical perspective, Tyers and Pienaar used WP to extract pairs of corresponding words in different languages [95]. Serrano et al, used WP corpus along with two others, to build up statistical models concerning fundamental concepts of patterns of word appearance in the text and vocabulary size [84]. Gabrilovich and Markovitch, introduced a method to calculate semantic relatedness of text fragments by extracting a “high-dimensional space of concepts” from WP [27]. In a recent paper [53] Kornai argued that the maturity of WP and the activity on it are important indicators for the chances of survival of a language in the digital age.

Our approach to WP as a text corpus is based on readability measures. We analyzed the readability of the English WP by a simple empirical formula suggested by Gunning [31,32] in the middle of last century for the English language:

$$F = 0.4 \left(\frac{\# \text{ of words}}{\# \text{ of sentences}} + 100 \frac{\# \text{ of complex words}}{\# \text{ of words}} \right). \quad (1)$$

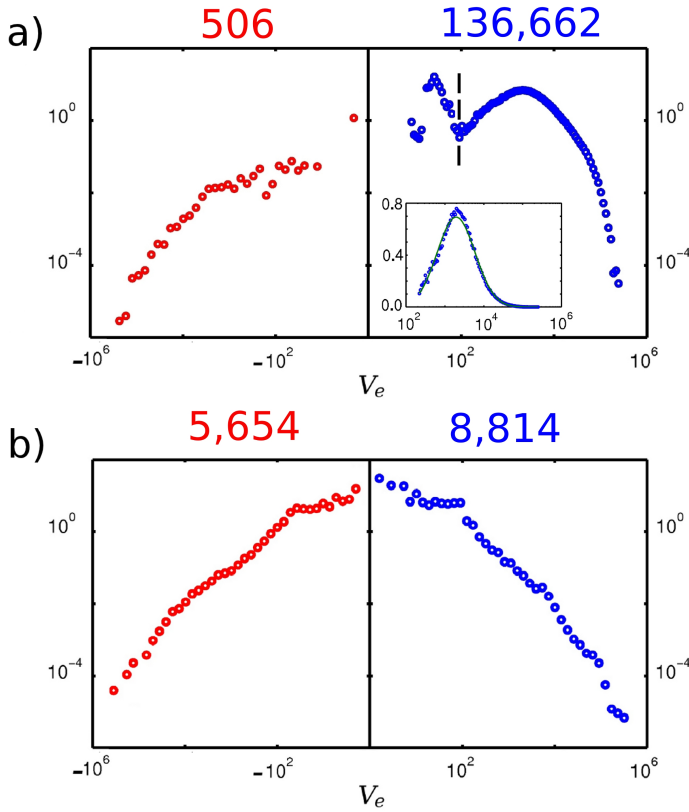


Fig. 5 Probability distribution function of edit volume (in bytes) for two typical editors: producer (a) and maintainer (b). Blue/red points representing the volume of added/deleted words in each edit. Total number of edits of each type are reported on top of each panel. The producer editor has much more adding edits, whereas the maintainer has comparable number of adding, deleting edits. The distribution of the volume of added parts by the producer editor has two peaks corresponding to volumes of few sentences and few paragraphs respectively (including the references and Wikimedia tags). *Inset* of panel (a) is a semi-logarithmic zoom on the right part of the added words volume distribution separated by the dashed line in the main panel. Fitted line to the inset is a log-normal distribution function.

where, complex words are those with three or larger syllables. The readability measure F is interpreted as the length of needed education time in years, to be able to read and understand the text.

We found out that the overall readability of English WP is high with $F = 15.8 \pm 0.4$ compared to other standard English corpora, for instance British National Corpus²⁷ with $F = 12.1 \pm 0.5$ [109]. However, readability is not homogeneous among articles in different topics. We observed that articles on more sophisticated topics or concepts, especially in science and philosophy are less readable than, e.g., biographical articles.

An interesting language edition of WP is “Simple Wikipedia”, which is meant to be a proper reference for those readers with weaker knowledge of English, e.g., children, language learners or non-native speakers. Editors of Simple are explicitly requested to use a simpler language, limited vocabulary, less complex words, shorter sentences, and easy structures.²⁸

In a recent work [9], Simple is examined by measuring the Flesch reading score [26] and it is found that Simple is not simple enough compared to other English texts, however with a positive trend in time towards more simplicity. There have been also attempts to use Simple WP for establishing text simplification algorithms [112,65,20], however with the assumption that Simple WP is really simple. The comparison of Simple and English WPs enables to study the ability of the editors to fulfill a preset task (namely enhance readability) and, at the same time, it also sheds more light to the concept of language complexity in general. We measured the readability index for a sample extracted from Simple WP [109]. We found it to be 10.8 ± 0.2 , i.e., indeed much lower than for the English WP but just as large as a corpus made of Wall Street Journal²⁹ articles.

To further analyze language complexity of Simple WP, we made the statistics at the word level, and surprisingly observed that vocabulary richness of Simple is comparable to that of main English

²⁷ <http://www.natcorp.ox.ac.uk>

²⁸ http://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages

²⁹ <http://www.wsj.com>

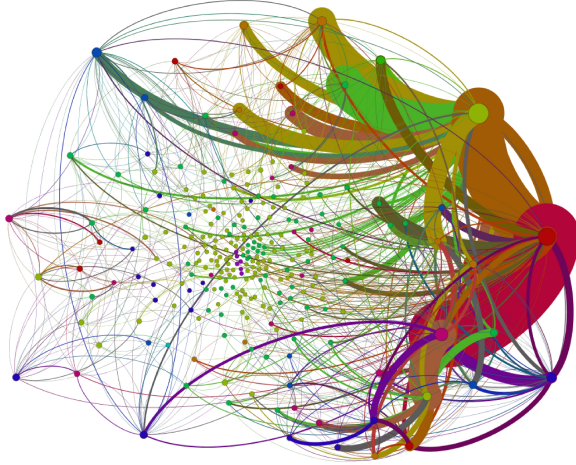


Fig. 6 A network representation of reverts in the history of the article on “Anarchism” in English WP. Nodes are editors and links are representing reverts. Size of the nodes is proportional to the total number of reverts, in which the editor is involved and width of the links is proportional to the number of total reverts between their corresponding pair of editors. Few nodes are strongly connected whereas most of the nodes are connected only through weak links. Completed triangles are clearly under-represented. The graph is generated Gephi, the open graph visualization platform, available at <http://gephi.org>.

WP. Moreover, by examining two fundamental laws of linguistics, namely Zipf’s law [114] and Herdan-Heaps’ law [37,36], we again confirmed that vocabulary richness of Simple and Main English WP are not significantly different [109], although the directives explicitly suggest self-restriction in this respect for Simple editors. Detailed analysis of longer units (n-grams) of words shows that the language of Simple is indeed less complex than that of Main but due to more frequent use of predefined language blocks, e.g., chains of words in the length of 4 or 5 words in Simple. Length of sentence is also shorter in Simple compared to Main. One can conclude that Simple editors solved fairly well the task to write more readable texts as compared to those in the main English WP without following slavishly the directives but mostly by reducing the variation of language compounds.

4.2 Conflicts and edit wars

In the process of creating a common product by various agents, occurrence of controversies due to different opinions are unavoidable. WP is neither a exception in this sense. WP editorial wars and disputes are known and studied phenomena [52,11,100,5,91,92,111]. Editorial wars could be evoked both by internal and external causes. For example life events of celebrities [111] or natural disasters [49] could conduct flows of editors to an article leading to tensions and disagreements. Apic et al. showed that disputes in WP are corresponding to real world geopolitical instabilities in many cases [5]. To study editorial wars in details, the first step is to establish an algorithm to locate and rank the debated articles among the relatively large number [91] of peacefully written articles. There are different proposals for this goal in the literature [52,11,100,92]. In the following Section we briefly describe our previously established method [92] for locating and ranking editorial wars.

4.2.1 Identification of controversial articles

The algorithm to identify disputed articles, introduced in [91,92,111], is based on counting mutual reverts by pairs of editors. Reverts are edits which bring the article exactly to a version before the last revision. In the other words, it is complete elimination of an editors single edit. And mutual reverting pairs are formed when editors revert each other at least once in each direction. In Fig. 6, a network representation of mutual reverts of the article on “Anarchism” in English WP is shown.

To eliminate the bias of vandalism and technical mistakes by unexperienced editors, a correction proportional to the “maturity” of reverting editor is applied by taking into account the number of edits she made. Finally, an extra factor, the number of editors involved is introduced leading controversy measure hereafter called M . Clearly, as the time goes on, more mutual reverts could happen in the history of the article. This makes M a dynamic, monotonically increasing variable. Having calculated M for all articles, we are able to find and rank most disputed ones and investigate them in details. We

carried out a detailed comparative study of possible single measures and found that M is in most cases as good as its alternatives if not better with the additional advantage of being applicable to different languages. The superiority of our single parameter measure was reinforced by a recent independent investigation [82].

Controversial topics: Based on the calculated controversy measure for articles in different languages, first conclusion is that, although there are sever editorial wars on some articles, but most of the articles in different languages evolve rather peacefully. However, the truly disputed articles consume a considerable amount of editorial resources. Interesting patterns are observed by comparing the debated titles in different language editions [108]. For instance, issues related to politics and religion are commonly among the most disputed articles in many language WPs, whereas, some category of topics only become controversial in specific languages. Science and philosophy in French and soccer clubs in Spanish WPs are examples of locally debated topics. There are even articles, in top of the controversy list in one WP, which is not even covered in other language edition, or does not have a separate article. Here examples are detailed articles around “Baha’i Faith” related topics in Persian WP. Finally, surprisingly, in the Hebrew WP, sport is debated as much as religion and politics.

4.2.2 Temporal features

The understanding of the emergence of conflicts, their escalation and resolution is important for maintaining WP and may give hints in general for techniques of conflict management. The controversy measure M enables the temporal analysis of editorial wars on short and long time scales.

Edit frequency: Intuitively more popular articles are subject to more collision of opinions and edit wars. However, the correlation between the average times between edits and the measure M is not significantly strong ($C = -0.03$).

Burstiness and conflict As discussed in Section 4.1.2, bursty trains of activity are clearly present on the editor level. Now, we focus on the edit trains of individual articles, i.e., for a particular article we consider the edits from all editors. We calculate the burstiness measure suggested in [29] as

$$B \equiv \frac{\sigma_\tau - m_\tau}{\sigma_\tau + m_\tau}, \quad (2)$$

where σ_τ and m_τ are the standard deviation and the average of inter-edit time intervals of each article. For a regular pattern with a delta function-like interval distribution, $B \rightarrow -1$, and for a random Poissonian process with a fixed event rate $B \rightarrow 0$. However, for fat-tailed distributions of time intervals, when the standard deviation diverged for infinite systems, $B \rightarrow 1$. The correlation between burstiness and controversy was also found to be rather small ($C=0.05$), considering the whole sample of articles in English WP. We also considered smaller samples of “featured” and “controversial” articles, based on the lists described in Section 2.1 and compared them to a sample of articles selected randomly. When looking at the distribution of the burstiness B we could not find significant differences between the three categories of articles. However, the B -values for the reverts, and especially for mutual reverts show significant increase of the burstiness, when going from featured through average to conflict articles (see Fig. 7).

To further investigate the temporal features of the editorial wars, we select two small samples of articles with the same average inter-edit time intervals, however one collected from largely disputed articles, and the other consisting of peaceful ones. Fig. 8 clearly shows that in the sample of disputed article, edits come in condense bursty trains with long range memories, whereas the statistics of the number of edits in the bursty periods in peaceful articles is much closer to a process. In a recently paper [46], it was claimed that edit time series can be described by a Poisson process, i.e., “edit-events are only short-term correlated”. Note, however, that due to the overwhelming number of peaceful articles a random mixture of articles does not sample conflicting ones. Therefore the conclusion of [46] is true for such articles but, as we demonstrated, not for controversial ones, where long-term memory is present.

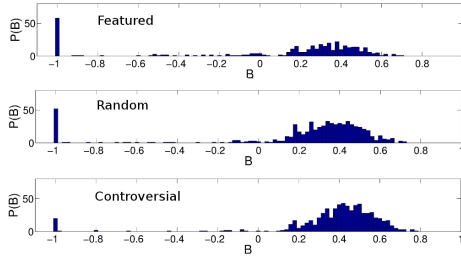


Fig. 7 Histogram of B values calculated only considering the temporal sequence of mutual reverts, for 3 different samples of articles; featured, randomly selected, and controversial articles. The last two samples are made based on the lists provided in Wikipedia (see Section 2.1).

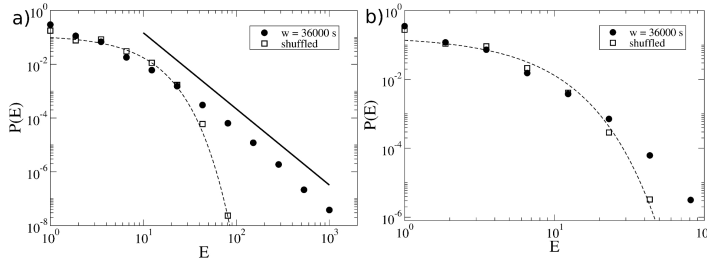


Fig. 8 Probability density function of the number of edits in bursty periods separated by a silence window of 10 hrs, for two samples of (a) highly controversial, and (b) peaceful articles with the same average inter-edit time interval. Black circles are the original data and empty squares are the shuffled sequence of the same intervals. Bursty periods with very large number of events are visible in (a), whereas the decay of the probability density function is very close to exponential in (b), indicating presence of memory effects in the case of controversial articles. *This figure is originally published in [111] under the terms of the Creative Commons Attribution License.*

4.2.3 Talk pages, conflict and coordination

As mentioned in Section 2.3, talk pages are channels to resolve the editorial disagreements in a more civilized manner than overriding each others edits and “talk before you type” is considered as the ideal mechanism of coordination in WP [97]. In a novel approach, Hautasaari and Ishida investigated the role of talk pages in coordination of translation of articles from English to Finnish, French, and Japanese [35]. They conclude that most of the debates in this field are about naming issues and not much about the content. Schneider et al. performed very detailed analysis of Talk pages from 100 articles manually and talk pages from 5000 articles quantitatively [81]. Their results for the category of controversial articles suggests “significant variance between discussion threads (different sub-topics in the talk page of a certain article) on their talk pages”, such that the distribution of the length of single threads is quite heterogeneous. Many threads are rather short, with few comments, and few of them become extremely long with numerous comments. This is in accordance with the results of [30], where a preferential attachment model to explain the discussion cascades in the talk pages was presented.

We measure the length of the talk pages for all articles. The correlation between talk page length and M for the English WP is much more significant ($C=0.54$) than that with the edit frequency. It indicates that most of the debates are reflected in talk pages simultaneous to edit wars directly on the article. This is partially supported in [45], where a method to detect “peaks” in talk pages is presented and showed that larger peaks mostly co-occur with peaks in editorial activities in a distance of 2 days. However, there are substantial differences between WPs on different languages in the usage of talk pages. In general, less developed WPs use talk pages less but even rather mature WPs, like the German one do not fight out controversies on the talk pages. (For a collection of visualizations and other related materials to edit wars, see <http://www.phy.bme.hu/>.)

Discussion Networks: In contrast to the revert network of editors, which can be constructed rather straight-forwardly, creation of talk page networks need more sophisticated algorithms. Laniado et al. constructed three types of talk networks by considering i) direct replies between users in article discussion pages, ii) direct replies in user talk pages, and iii) personal messages posted on the talk page of another user [56]. The conclusion of this studies suggests the presence of dissortativity in outgoing links and assortativity in incoming links.

Our case study of the talk page of “Safavid dynasty” showed that most of the comments are exchanged between few editors, who are actively editing the articles. In addition, the occurrence of clusters is very rare, such that most of the conversations are between pairs of editors and not bigger groups of them. This is in accord to analysis on mutual reverts which shows only few editors are responsible for large amount of edit wars [111]. By fine investigation on those few user-names very active in controversial articles, we could recall many of them from our list of “Bad Editors” introduced in Section 4.1.1.

Language complexity and sentiment: Laniado et al, studied the emotional aspects of talk page discussions by measuring sentiment of the comments and found “replies are on average more positive than the comments they reply to, and editors having similar emotional styles are more likely to interact with each other.” Moreover, they found that editors with more social power, i.e. admins, talk more positively and interestingly this is also the case for female editors [55].

We measured the readability of talk pages based on Eq. 1 and compared it to the readability of articles for two samples of controversial and peaceful articles. In both cases there is a significant reduction in readability, going from articles to corresponding talk pages [109]. However, the reduction is much more significant for the controversial articles. This can be explained by previous sociological theories on the effect of destructive conflict on complexity reduction of language [80]; In simple words, when people talk with more temper, they use less sophisticated language.

4.2.4 Leader-follower behavior in conflict

The community of editors is structured though it is not easy to unfold it. When studying the talk pages of highly edited articles, it becomes clear that editorial behavior is influenced beyond the content also by personal relationships [79]. There are dominant editors and others, who only follow them. Such relationships largely influence the emerging editor network. The easiest way to detect related behavior is to concentrate on leader-follower pairs. These are pairs of editors (say, A and B), who often act in a specific order, i.e., A always precedes B within a reasonable time, e.g., 1 day. As we are interested in the difference between peaceful and conflict articles, we concentrated on the leader-follower phenomenon in reverts [79]. We defined the following process as an event: A reverts C and (within one day) B reverts C , where C is fixed only for this specific event. Confining our interest to reverts restricted considerably the statistics, however, here significant differences between the two groups of articles could be observed.

We took two different edit history samples of WP. The first sample consisted all reverts of the 837 articles with M value above 10^5 (conflict articles) ordered by time. In order to avoid the effects of vandalism, we excluded reverter-reverted editor pairs consisting at least of one IP address or bot. Moreover, to gain a better focus on leader-follower relationships and not the effects produced solely by editorial wars between two editors, we also excluded repeated reverts where the reverter-reverted pair was the same and no other reverts happened between these two reverts. This seed consisted of 303397 reverts. We took a sample of 12470 articles with M value under 500 (peaceful articles), where the number of reverts was approximately the same. We also created randomized versions of these samples.

The results are summarized in Fig. 9 (a), from which it is clear that conflict articles have an enhanced amount of leader-follower patterns. This is expected as in this case parties are formed, where the hierarchy of editors can manifest itself. We see that even sequences of length $l = 10$ occur. Fig. 9 (b) indicates that leader-follower actions have a characteristic time T of about 2 minutes. This is a surprisingly short period and underlines the personal rather than contextual motivation.

4.2.5 War scenarios

The characterization of the temporal evolution of conflicts is crucial for their typology and understanding. Our measure M is particularly suitable for such a study. We investigated controversial articles of English WP from this point of view. Instead of the real time we use the number of edits as a control parameter. This way we eliminate several sources of temporal inhomogeneities like maturing the whole WP, differences in the sizes of the articles, and external events motivating editors to focus on an article [75, 74].

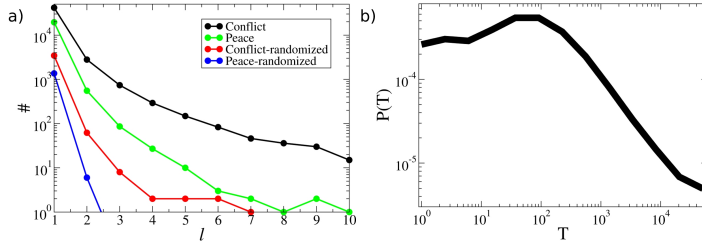


Fig. 9 Leader-follower statistics from revert series of peaceful and conflict articles as well as their randomized versions. a) Statistics of leader-follower sequences. b) Probability distribution function of elapsed time between leader and follower events.

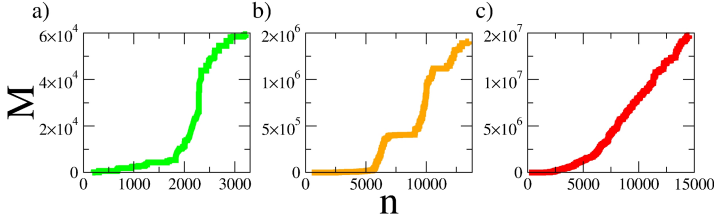


Fig. 10 Evolution of the controversy measure M as a function of number of edits on the articles n , for a) “Bombing of Dresden in World War II”, b) “Japan”, and c) “Anarchism”. In (a), after a single period of editorial war, the article reaches the stable consensus, whereas in (b) temporary consensus states are altered by new conflict periods. In (c) the rate of initiation of new conflicts is very high, such that no consensus is ever reached.

Three different scenarios of wars could be distinguished [111] from the temporal evolution of M (Fig. 10):

i) Consensus after war, Fig. 10 (a); After a smooth initial increase of M , an intense period of war appears and once the conflict is resolved, the article reaches consensus and farther edits are mostly on polishing and improving the presentation quality. This is the scenario for most of the disputed articles in English WP [111]. ii) Stepwise conflicts, Fig. 10 (b); After the first cycle of conflict-resolution, the consensus state might be altered mainly because of one of two reasons, namely occurrence of an external event which generates new controversy or arrival of new editors, who are not satisfied with the previously compromised content of the article. Therefore, other conflict-resolution cycles may appear in the overall history of the article. iii) Never-ending war, Fig. 10 (c); If the rate of incoming editors or external events related to the topic of the article, is considerably larger than the typical time to reach consensus, even a temporary equilibrium cannot be achieved and the increase of M becomes permanent. This is the case of highly popular and live-object articles. Number of such articles in English WP does not exceed few hundreds (compared to some millions, the total number of articles).

4.2.6 Agent-based modeling

Motivated by empirical results on editorial wars in WP, we aimed at providing a minimalistic agent-based model capturing the main features of the wars [94]. The model belongs to the class of bounded confidence models of opinion dynamics introduced by Deffuant et al. [22]. It consists of two types of elements; N_e editors and one article. In each Monte Carlo step, editors interact if their scalar opinions $x_i \in [0, 1]$, $i = 1 \dots N_e$ are already closer to each other than a threshold value ϵ_T and then they adopt the opinion of the arithmetic mean. An editor edits the article if she finds it in a state $A \in [0, 1]$ with a difference larger than ϵ_A to x_i , otherwise she revises her own opinion which gets closer to the article state by an amount controlled by a parameter μ_A . In addition, editors can be replaced in each step by new ones with a constant rate p_{new} .

Fixed editorial pool: To evaluate the outcome of the model, initially p_{new} is set to 0, which leads to consensus for the whole parameter space, meaning that after sufficiently time A becomes constant. However, the relaxation time to consensus very much depends on the parameters set. There are three different scenarios to approach the consensus state: i) for small values of μ_A , system needs astronomically long time to reach the final state, although A is always very close to the system average of x_i . ii) Intermediate values of μ_A puts the system into an oscillatory phase, in which A fluctuates largely between two extreme values, however ending up with one of them in a relatively shorter time. iii) Large

values of μ_A leads to exaggerated fluctuations of A , however with fast convergence of extremist editors and a shorter relaxation time compare to the previous cases.

Dynamic editorial pool: The constant rate of replacement of old editors by new ones with random opinions can hinder the system to reach a time-independent state. The interplay of two time scales, namely relaxation and renewal leads to three different phases. For small renewal rate, the system experiences well separated periods of conflict (large fluctuation of A) followed by long consensus state, with only minor fluctuations of A , whereas for larger rates, even temporary consensus state is never reached and the system is constantly pushed outwards consensus. Finally, there is a narrow transition regime in the phase space, in which there are numerous short periods of peace and war appearing consequently. This three regimes are corresponding to the war scenarios discussed in Section. 4.2.5. In order to make the comparison to real data more feasible, we defined the cumulative amount of conflict in the system as the total sum of changes in the position of the article up to time t ($t \times N$ pairs of editor-editor and editor-article interactions)

$$S(t) = \sum_{i=1}^t \sum_{j=1}^N |A(i) - A(i-1)|. \quad (3)$$

In fact, the temporal evolution of S shows similar patterns as that of M . For low renewal rate we have a conflict period followed by a peaceful one, where only minor changes happen. At large renewal rate we have permanent war. In between there is an alternation of conflict and peaceful regions. These results support the intuitive picture that increasing the the large rate of newcomer editors increase the vulnerability of the consensus.

5 Conclusions

In this paper we surveyed recent work on WP and extended it by some new results. Our studies covered multilingual aspects and focused on the mechanisms and consequences of collaborative value production. The analysis of daily and weekly patterns of the editorial activity made it possible to identify the contributions from different parts of the world to such globally edited WPs as the English, the Spanish or the Arabic as well as to point out cultural differences in editing habits. The "wisdom of the crowd" seems to cope better with some tasks than pre-designed directives as the case of Simple WP demonstrates. Our main focus was to characterize and understand how conflicts emerge and get eventually resolved. While most of the WP articles are edited in a peaceful, constructive atmosphere, some of the most popular articles are rather controversial. In order to be able to study the conflict pages systematically, we developed a simple measure to identify them automatically. We have found interesting differences between peaceful and conflict pages in their dynamics as the edit activity of the latter is a long range correlated process in contrast to that of average (peaceful) pages. The language of the talk pages of conflict articles gets more reduced in complexity than that of regular articles and the leader-follower behavior is more intensive. The temporal evolution of the measure M enabled to distinguish between different types of conflicts (single conflict with resolution, multiple conflicts, permanent war). Finally, we showed that simple multi-agent modeling based on opinion dynamics can reproduce some of our findings.

Acknowledgements We would like to thank our collaborators: Gerardo Iñiguez, Kimmo Kaski, András Kornai, András Rung, Maxi San Miguel, Róbert Sumi, János Török. Discussions, advise and help with the data are gratefully acknowledged to Farzaneh Kaveh, Santo Fortunato, Márton Mestyán, Andrzej Nowak, Hoda Sepehri Rad, Attila Zséder, Gábor Recski, Peter Reuvern, and Katarzyna Samson.

References

1. Aaltonen, A., Lanzara, G.F.: Governing complex social production in the internet: The emergence of a collective capability in Wikipedia (2011). In Decade in Internet Time symposium
2. Adler, B.T., de Alfaro, L.: A content-driven reputation system for the Wikipedia. Tech. Rep. ucsc-crl-06-18, School of Engineering, University of California, Santa Cruz (2006)

3. Adler, B.T., de Alfaro, L., Mola-Velasco, S., Rosso, P., West, A.: Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In: A. Gelbukh (ed.) *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*, vol. 6609, pp. 277–288. Springer Berlin / Heidelberg (2011)
4. Almeida, R.B., Mozafari, B., Cho, J.: On the evolution of wikipedia. In: *Proceedings of the International Conference on Weblogs and Social Media, ICWSM'07* (2007)
5. Apic, G., Betts, M.J., Russell, R.: Content disputes in Wikipedia reflect geopolitical instabilit. *PLoS ONE* **06**(6), e20,902 (2011)
6. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: K. Aberer, K.S. Choi, N. Noy, D. Allemang, K.I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, P. Cudr-Mauroux (eds.) *The Semantic Web, Lecture Notes in Computer Science*, vol. 4825, pp. 722–735. Springer Berlin / Heidelberg (2007)
7. Ayers, P., Priedhorsky, R.: Wikilit: collecting the wiki and wikipedia literature. In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration, WikiSym '11*, pp. 229–230. ACM, New York, NY, USA (2011)
8. Barabási, A.L.: The origin of bursts and heavy tails in human dynamics. *Nature* **435**(7039), 207–211 (2005)
9. Besten, M.D., Dalle, J.: Keep it simple: A companion for Simple Wikipedia? *Industry & Innovation* **15**(2), 169–178 (2008)
10. Bohannon, J.: Tracking people's electronic footprints. *Science* **314**(5801), 914–916 (2006)
11. Brandes, U., Lerner, J.: Visual analysis of controversy in user-generated encyclopedias. *Information Visualization* **7**(1), 34–48 (2008)
12. Buriol, L.S., Castillo, C., Donato, D., Leonardi, S., Millozzi, S.: Temporal analysis of the wikigraph. In: *In Proc. of Web Intelligence, Hong Kong*, pp. 45–51 (2006)
13. Butler, B., Joyce, E., Pike, J.: Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In: *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, CHI '08*, pp. 1101–1110. ACM, New York, NY, USA (2008)
14. Capocci, A., Rao, F., Caldarelli, G.: Taxonomy and clustering in collaborative systems: The case of the on-line encyclopedia Wikipedia. *EPL (Europhysics Letters)* **81**(2), 28,006 (2008)
15. Capocci, A., Servedio, V.D.P., Colaiori, F., Buriol, L.S., Donato, D., Leonardi, S., Caldarelli, G.: Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Phys. Rev. E* **74**, 036,116 (2006)
16. Zlatić, V., Štefančić, H.: Model of Wikipedia growth based on information exchange via reciprocal arcs. *EPL (Europhysics Letters)* **93**(5), 58,005 (2011)
17. Zlatić, V., Božičević, M., Štefančić, H., Domazet, M.: Wikipedias: Collaborative web-based encyclopedias as complex networks. *Phys. Rev. E* **74**, 016,115 (2006)
18. Chakrabarti, B.K., Chakraborti, A., Chatterjee, A. (eds.): *Econophysics and Sociophysics: Trends and Perspectives*. Wiley-VCH Verlag GmbH & Co, Berlin (2006)
19. Cohen, J.: Computational methods for historical research on Wikipedia's archives. *e-Research* **1**(2), 67–72 (2010)
20. Coster, W., Kauchak, D.: Simple English Wikipedia: a new text simplification task. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, HLT '11*, pp. 665–669. Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
21. Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., Kleinberg, J.: Echoes of power: Language effects and power differences in social interaction. to appear in proceeding of WWW'12, priprint; arXiv:1112.3670 (2012)
22. Deffuant, G., Neau, D., Amblard, F., Weisbuch, G.: Mixing beliefs among interacting agents. *Adv. Complex Syst.* **3**(4), 87–98 (2000)
23. Denoyer, L., Gallinari, P.: *Acm sigir forum*. Eueopean Academy of Management Annual Conference 2010, Rome, Italy. **40**(1) (2006)
24. Derthick, K., Tsao, P., Kriplean, T., Borning, A., Zachry, M., McDonald, D.: Collaborative sensemaking during admin permission granting in Wikipedia. In: A. Ozok, P. Zaphiris (eds.) *Online Communities and Social Computing, Lecture Notes in Computer Science*, vol. 6778, pp. 100–109. Springer Berlin / Heidelberg (2011)
25. F., O.: Wikipedia. a quantitative analysis. Ph.D. thesis, University Rey Juan Carlos, Madrid, Spain (2009)
26. Flesch, R.: *How to Write Plain English*. Harper and Row, New York (1979)
27. Gabrilovich, E., Markovitch, S.: Wikipedia-based semantic interpretation for natural language processing. *J. Artif. Int. Res.* **34**, 443–498 (2009)
28. Giles, J.: Internet encyclopaedias go head to head. *Nature* **438**, 900 (2005)
29. Goh, K.I., Barabási, A.L.: Burstiness and memory in complex systems. *EPL* **81**(4), 48,002 (2008)
30. Gómez, V., Kappen, H.J., Kaltenbrunner, A.: Modeling the structure and evolution of discussion cascades. In: *Proceedings of the 22nd ACM conference on Hypertext and hypermedia, HT '11*, pp. 181–190. ACM, New York, NY, USA (2011)
31. Gunning, R.: *The technique of clear writing*. NY: McGraw-Hill International Book Co., New York (1952)
32. Gunning, R.: The fog index after twenty years. *Journal of Business Communication* **6**(2), 3–13 (1969)
33. Halavais, A., Lackaff, D.: An analysis of topical coverage of Wikipedia. *Journal of Computer-Mediated Communication* **13**(2), 429–440 (2008)
34. Hardy, D., Frew, J., Goodchild, M.F.: Volunteered geographic information production as a spatial process. *International Journal of Geographical Information Science* **26**(7), 1191–1212 (2012)
35. Hautasaari, A., Ishida, T.: Analysis of discussion contributions in translated Wikipedia articles. In: *Proceedings of the 4th international conference on Intercultural Collaboration, ICIC '12*, pp. 57–66. ACM, New York, NY, USA (2012)

36. Heaps, H.S.: *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc., Orlando, FL, USA (1978)
37. Herdan, G.: *Quantitative linguistics*. Butterworths, Washington (1964)
38. Holloway, T., Bozicevic, M., Brner, K.: Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity* **12**(3), 30–40 (2007)
39. Hu, M., Lim, E.P., Sun, A., Lauw, H.W., Vuong, B.Q.: Measuring article quality in Wikipedia: models and evaluation. In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM '07*, pp. 243–252. ACM, New York, NY, USA (2007)
40. Javanmardi, S., Ganjisaffar, Y., Lopes, C., Baldi, P.: User contribution and trust in wikipedia. In: *Collaborative Computing: Networking, Applications and Worksharing, 2009. CollaborateCom 2009. 5th International Conference on*, pp. 1–6 (2009)
41. Javanmardi, S., Lopes, C.: Statistical measure of quality in Wikipedia. In: *Proceedings of the First Workshop on Social Media Analytics, SOMA '10*, pp. 132–138. ACM, New York, NY, USA (2010)
42. Javanmardi, S., Lopes, C., Baldi, P.: Modeling user reputation in Wikis. *Statistical Analysis and Data Mining* **3**(2), 126–139 (2010)
43. Jones, J.: Patterns of revision in online writing. *Written Communication* **25**(2), 262–289 (2008)
44. Jullien, N.: What we know about Wikipedia: A review of the literature analyzing the project(s) (2012). Available at SSRN: <http://ssrn.com/abstract=2053597>
45. Kaltenbrunner, A., Laniado, D.: There is no deadline - time evolution of Wikipedia discussions. In: *Proceedings of the 8th International Symposium on Wikis and Open Collaboration, WikiSym'12*. Linz (2012)
46. Kämpf, M., Tismer, S., Kantelhardt, J.W., Muchnik, L.: Fluctuations in Wikipedia access-rate and edit-event data. *Physica A: Statistical Mechanics and its Applications* (2012)
47. Karkulahti, O., Kangasharju, J.: Surveying Wikipedia activity: Collaboration, commercialism, and culture. In: *Information Networking (ICOIN), 2012 International Conference on*, pp. 384–389 (2012)
48. Karsai, M., Kaski, K., Barabási, A.L., Kertész, J.: Universal features of correlated bursty behaviour. *Sci. Rep.* **2** (2012)
49. Keegan, B., Gergle, D., Contractor, N.: Hot off the wiki: dynamics, practices, and structures in Wikipedia's coverage of the Tōhoku catastrophes. In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration, WikiSym '11*, pp. 105–113. ACM, New York, NY, USA (2011)
50. Kittur, A., Chi, E.H., Suh, B.: What's in Wikipedia?: mapping topics and conflict using socially annotated category structure. In: *Proceedings of the 27th international conference on Human factors in computing systems, CHI '09*, pp. 1509–1512. ACM, New York, NY, USA (2009)
51. Kittur, A., Kraut, R.E.: Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In: *Proceedings of the 2008 ACM conference on Computer supported cooperative work, CSCW '08*, pp. 37–46. ACM, New York, NY, USA (2008)
52. Kittur, A., Pendleton, B.A., Suh, B., Mytkowicz, T.: Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In: *CHI 07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2007)
53. Kornai, A.: *Language death in the digital age*. to be published (2012)
54. Lam, S.T.K., Uduwage, A., Dong, Z., Sen, S., Musicant, D.R., Terveen, L., Riedl, J.: Wp:clubhouse?: an exploration of Wikipedia's gender imbalance. In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration, WikiSym '11*, pp. 1–10. ACM, New York, NY, USA (2011)
55. Laniado, D., Castillo, C., Kaltenbrunner, A., Fuster Morell, M.: Emotions and dialogue in a peer-production community: the case of Wikipedia. In: *Proceedings of the 8th International Symposium on Wikis and Open Collaboration, WikiSym'12*. Linz (2012)
56. Laniado, D., Tasso, R., Volkovich, Y., Kaltenbrunner, A.: When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. In: *5th International AAAI Conference on Weblogs and Social Media, ICWSM 2011*, pp. 177–184 (2011)
57. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Van Alstyne, M.: Computational social science. *Science* **323**(5915), 721–723 (2009)
58. Lee, J.B., Cabunducan, G., Cabarle, F.G.C., Castillo, R., Malinao, J.A.: Uncovering the social dynamics of online elections **18**(4), 487–505 (2012)
59. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Governance in social media: A case study of the Wikipedia promotion process. In: *Proceedings of the International Conference on Weblogs and Social Media, ICWSM'10* (2010)
60. Leuf, B., Cunningham, W.: *The Wiki way: quick collaboration on the Web*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA (2001)
61. Luyt, B., Aaron, T.C.H., Thian, L.H., Hong, C.K.: Improving Wikipedia's accuracy: Is edit age a solution? *J. Am. Soc. Inf. Sci. Technol.* **59**(2), 318–330 (2008)
62. Massa, P.: Social networks of Wikipedia. In: *Proceedings of the 22nd ACM conference on Hypertext and hypermedia, HT '11*, pp. 221–230. ACM, New York, NY, USA (2011)
63. Masucci, A.P., Kalampokis, A., Eguíluz, V.M., Hernández-García, E.: Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. *PLoS ONE* **6**(2), e17,333 (2011)
64. Muchnik, L., Itzhack, R., Solomon, S., Louzoun, Y.: Self-emergence of knowledge trees: Extraction of the Wikipedia hierarchies. *Phys. Rev. E* **76**, 016,106 (2007)
65. Napoles, C., Dredze, M.: Learning simple Wikipedia: a cogitation in ascertaining abecedarian language. In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing, CL&W '10*, pp. 42–50. Association for Computational Linguistics, Stroudsburg, PA, USA (2010)

66. Nielsen, F.A.: Wikipedia research and tools: Review and comments (2011). Available at http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/6012/pdf/imm6012.pdf
67. Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F.A., Lanamäki, A.: The peoples encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia (2012). Available at SSRN: <http://ssrn.com/abstract=2021326>
68. Ortega, F., Gonzalez-Barahona, J., Robles, G.: On the inequality of contributions to Wikipedia. In: Hawaii International Conference on System Sciences, Proceedings of the 41st Annual, p. 304 (2008)
69. Park, T.K.: The visibility of Wikipedia in scholarly publications. *First Monday* **16**(8) (2011)
70. Pentzold, C., Seidenglanz, S.: Foucault@wiki: first steps towards a conceptual framework for the analysis of wiki discourses. In: Proceedings of the 2006 international symposium on Wikis, WikiSym '06, pp. 59–68. ACM, New York, NY, USA (2006)
71. Ponzetto, S.P., Strube, M.: Knowledge derived from Wikipedia for computing semantic relatedness. *J. Artif. Int. Res.* **30**, 181–212 (2007)
72. Potthast, M., Stein, B., Gerling, R.: Automatic vandalism detection in Wikipedia. In: Proceedings of the IR research, 30th European conference on Advances in information retrieval, ECIR'08, pp. 663–668. Springer-Verlag, Berlin, Heidelberg (2008)
73. Ratkiewicz, J., Flammini, A., Menczer, F.: Traffic in social media i: Paths through information networks. In: Social Computing (SocialCom), 2010 IEEE Second International Conference on, pp. 452–458 (2010)
74. Ratkiewicz, J., Fortunato, S., Flammini, A., Menczer, F., Vespignani, A.: Characterizing and modeling the dynamics of online popularity. *Phys. Rev. Lett.* **105**, 158,701 (2010)
75. Ratkiewicz, J., Menczer, F., Fortunato, S., Flammini, A., Vespignani, A.: Traffic in social media ii: Modeling bursty popularity. In: Social Computing (SocialCom), 2010 IEEE Second International Conference on, pp. 393–400 (2010)
76. Reinoso, A.J., González-Barahona, J.M., Muñoz Mansilla, R., Herráiz Tabernero, I.: Temporal characterization of the requests to Wikipedia. In: Proceedings of the 5th International Workshop on New Challenges in Distributed Information Filtering and Retrieval (DART 2011), vol. 771 (2011)
77. Restivo, M., van de Rijt, A.: Experimental study of informal rewards in peer production. *PLoS ONE* **07**(7), e34,358 (2012)
78. Roth, C., Taraborelli, D., Gilbert, N.: Measuring wiki viability: an empirical assessment of the social dynamics of a large sample of wikis. In: Proceedings of the 4th International Symposium on Wikis, WikiSym '08, pp. 27:1–27:5. ACM, New York, NY, USA (2008)
79. Rung, A., Yasseri, T., Kornai, A., Kertész, J.: Editorial relations in controversial Wikipedia articles. to be published (2012)
80. Samson, K., Nowak, A.: Linguistic signs of destructive and constructive processes in conflict. *IACM 23rd Annual Conference Paper* (2010)
81. Schneider, J., Passant, A., Breslin, J.: A qualitative and quantitative analysis of how Wikipedia talk pages are used. In: Proceedings of the WebSci10: Extending the Frontiers of Society, April 26–27th, 2010, Raleigh, NC: US., pp. 1–7 (2010)
82. Sepehri Rad, H., Barbosa, D.: Identifying controversial articles in Wikipedia: A comparative study. In: Proceedings of the 8th International Symposium on Wikis and Open Collaboration, WikiSym'12. Linz (2012)
83. Sepehri Rad, H., Makazhanov, A., Rafiei, D., Barbosa, D.: Leveraging editor collaboration patterns in Wikipedia. In: Proceedings of the 23rd ACM conference on Hypertext and social media, HT '12, pp. 13–22. ACM, New York, NY, USA (2012)
84. Serrano, M.A., Flammini, A., Menczer, F.: Modeling statistical properties of written text. *PLoS ONE* **4**(4), e5372 (2009)
85. Silva, F., Viana, M., Travenolo, B., da F. Costa, L.: Investigating relationships within and between category networks in Wikipedia. *Journal of Informetrics* **5**(3), 431–438 (2011)
86. Smets, K., Goethals, B., Verdonk, B.: Automatic vandalism detection in Wikipedia: towards a machine learning approach. In: AAAI Workshop Wikipedia and Artificial Intelligence: an Evolving Synergy, WikiAI08, pp. 43–48. Association for the Advancement of Artificial Intelligence (2008)
87. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using Wikipedia. In: proceedings of the 21st national conference on Artificial intelligence, vol. 2, pp. 1419–1424. AAAI Press (2006)
88. Stvilia, B., Twidale, M.B., Smith, L.C., Gasser, L.: Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology* **59**(6), 983–1001 (2008)
89. Suchecki, K., Salah, A., Gao, C., Scharnhorst, A.: Evolution of Wikipedia's category structure. *Advances in Complex Systems* **15**(supp01), 1250,068 (2012)
90. Suh, B., Convertino, G., Chi, E.H., Pirolli, P.: The singularity is not near: slowing growth of wikipedia. In: Proceedings of the 5th International Symposium on Wikis and Open Collaboration, WikiSym '09, pp. 8:1–8:10. ACM, New York, NY, USA (2009)
91. Sumi, R., Yasseri, T., Rung, A., Kornai, A., Kertész, J.: Characterization and prediction of Wikipedia edit wars. In: Proceedings of the ACM WebSci'11, Koblenz, Germany, pp. 1–3 (2011)
92. Sumi, R., Yasseri, T., Rung, A., Kornai, A., Kertész, J.: Edit wars in Wikipedia. In: Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), pp. 724–727 (2011)
93. Taraborelli, D., Ciampaglia, G.: Beyond notability. collective deliberation on content inclusion in Wikipedia. In: Self-Adaptive and Self-Organizing Systems Workshop (SASOW), 2010 Fourth IEEE International Conference on, pp. 122–125 (2010)
94. Török, J., Iniguez, G., Yasseri, T., San Miguel, M., Kaski, K., Kertész, J.: Opinions, conflicts and consensus: Modeling social dynamics in a collaborative environment. submitted, preprint; arXiv:1207.4914 (2012)

95. Tyers, F., Pienaar, J.: Extracting bilingual word pairs from Wikipedia. In: Proceedings of the SALT MIL Workshop at Language Resources and Evaluation Conference, LREC08 (2008)
96. Ung, H.M., Dalle, J.M.: Characterizing online communities with their signals. Eueopean Academy of Management Annual Conference 2010, Rome, Italy. (2010)
97. Viegas, F.B., Wattenberg, M., Kriss, J., van Ham, F.: Talk before you type: Coordination in Wikipedia. In: System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on, p. 78 (2007)
98. Völkel, M., Kröttsch, M., Vrandečić, D., Haller, H., Studer, R.: Semantic wikipedia. In: Proceedings of the 15th international conference on World Wide Web, WWW '06, pp. 585–594. ACM, New York, NY, USA (2006)
99. Voss, J.: Measuring Wikipedia (2005). International Conference of the International Society for Scientometrics and Informetrics : 10th, Stockholm (Sweden), 24–28 July 2005
100. Vuong, B.Q., Lim, E.P., Sun, A., Le, M.T., Lauw, H.W., Chang, K.: On ranking controversies in Wikipedia: models and evaluation. In: Proceedings of the international conference on Web search and web data mining, WSDM '08, pp. 171–182. ACM, New York, NY, USA (2008)
101. Wattenberg, M., Viégas, F., Hollenbach, K.: Visualizing activity on wikipedia with chromograms. In: C. Baranauskas, P. Palanque, J. Abascal, S. Barbosa (eds.) Human-Computer Interaction INTERACT 2007, *Lecture Notes in Computer Science*, vol. 4663, pp. 272–287. Springer Berlin / Heidelberg (2007)
102. West, A.G., Kannan, S., Lee, I.: Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata. In: Proceedings of the Third European Workshop on System Security, pp. 22–28
103. Wikipedia: World wide web — wikipedia, the free encyclopedia (2012). URL http://en.wikipedia.org/w/index.php?title=World_Wide_Web&oldid=508583126. [Online; accessed 22-August-2012]
104. Wilkinson, D.M.: Strong regularities in online peer production. In: Proceedings of the 9th ACM conference on Electronic commerce, EC '08, pp. 302–309. ACM, New York, NY, USA (2008)
105. Wilkinson, D.M., Huberman, B.A.: Assessing the value of cooperation in Wikipedia. *First Monday* **12**(4) (2007)
106. Wu, G., Harrigan, M., Cunningham, P.: Characterizing Wikipedia pages using edit network motif profiles. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents, SMUC '11, pp. 45–52. ACM, New York, NY, USA (2011)
107. Wu, Q., Irani, D., Pu, C., Ramaswamy, L.: Elusive vandalism detection in Wikipedia: a text stability-based approach. In: Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM '10, pp. 1797–1800. ACM, New York, NY, USA (2010)
108. Yasseri, T., et. al.: Most controversial topics in wikipedia: A multilingual analysis. In preparation (2012)
109. Yasseri, T., Kornai, A., Kertész, J.: A practical approach to language complexity: a Wikipedia case study. to appear in PLoS ONE, preprint; arXiv:1204.2765 (2012)
110. Yasseri, T., Sumi, R., Kertész, J.: Circadian patterns of Wikipedia editorial activity: A demographic analysis. *PLoS ONE* **7**(1), e30,091 (2012)
111. Yasseri, T., Sumi, R., Rung, A., Kornai, A., Kertész, J.: Dynamics of conflicts in Wikipedia. *PloS ONE* **7**(6), e38,869 (2012)
112. Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., Lee, L.: For the sake of simplicity: unsupervised extraction of lexical simplifications from Wikipedia. In: Human Language Technologies 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, pp. 365–368. Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
113. Zesch, T., Müller, C., Gurevych, I.: Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In: Proceedings of the Conference on Language Resources and Evaluation, LREC (2008)
114. Zipf, G.K.: The psycho-biology of language: an introduction to dynamic philology. The MIT Press, Cambridge, MA (1935)